African Journal of Hospitality, Tourism and Leisure. ISSN: 2223-814X Research Note

December 2022, Vol 11, 6, pp.1913-1923

Predicting the Number of Tourists in-Flow to Kenya Using Seasonal Autoregressive Integrated Moving Average Model

Dennis Obiri Gechore
Department of Mathematics & Statistics, University of Embu, Embu, Kenya

Edwin Benson Atitwa
Department of Mathematics & Statistics, University of Embu, Embu, Kenya

Patrick Kimani
Department of Mathematics & Statistics, University of Embu, Embu, Kenya

Maurice Wanyonyi*
Department of Mathematics & Statistics, University of Embu, Embu, Kenya, Email, wanyonyi0001@gmail.com

Corresponding Author

How to cite this article: Gechore, D.O., Atitwa, E.B., Kimani, P. & Wanyonyi, M. (2022). Predicting the Number of Tourists in-Flow to Kenya Using Seasonal Autoregressive Integrated Moving Average Model. African Journal of Hospitality, Tourism and Leisure, 11(6):1913-1923. DOI: **https://doi.org/10.46222/ajhtl.19770720.332**

Abstract

Tourism is the leading source of revenue to the Kenyan Government, contributing about 8.8% to the Kenya's Gross Domestic Product. Based on the 2019 report released by the ministry of tourism and wildlife, tourism industry contributed approximately $7.9 billion to the Kenya's budget. This study was therefore developed to predict the future numbers of tourists that will visit Kenya between 2023 and 2025. The Seasonal Autoregressive Integrated Moving Average time series model was applied for the prediction. The study used secondary data collected from the Ministry of Tourism and Wildlife. The data covered a period of 11 years from 2011 to 2022. The model was fitted to the real tourists' data using the time series algorithm implemented in R statistical software. Based on the Akaike Information Criterion, the ARIMA$(2,1,1)(0,1,0)_{12}$ was identified as the perfect model with minimum errors. The model passed the diagnostic test performed. Importantly, 95% confidence level prediction done for 3 years (2023-2025) using the model showed that the number of tourists expected to visit Kenya will increase significantly. Therefore, the study recommended that recreational facilities and accommodations should be maintained to cater for the high projected numbers of tourists. The study also recommended that the government of Kenya should strategize on how to beef up security to curb terrorism attacks and tribal conflicts which might discourage tourists.

Keywords: Time series model; prediction; SARIMA application; Kenya tourists forecasting; Akaike information criterion; R statistical software.

Introduction

African continent has the largest tourist attraction site in the world (Msofe & Mbago, 2019). Kenya being one of the leading tourist attraction country in Africa, it has a variety of tourist sites like the fort Jesus, flamingos at lake Nakuru, snow at the peak of Mount Kenya, Waterfalls, wildlife and museums among many more (Akuno et al., 2015). Over the decades, tourism has been the major sources of revenue to Kenya contributing about 8.8% to the country's Gross Domestic Product (GDP). As of 2019 report by the ministry of tourism and wildlife, tourism industry contributed approximately $7.9 billion to the Kenya's GDP (Makau et al., 2018). In the recent years, tourism industry has faced various insecurity challenges, including the Al-Shabaab attacks and the intertribal conflicts (Akuno et al., 2015). These

insecurity cases scare away tourists. Currently, the country is experiencing COVID-19 pandemic which has been a world-wide disaster. The pandemic has in addition led to a further decline in the number of tourists visiting Kenya due to lockdowns imposed in many countries globally in containment of its spread. These containment measures have made the country to lose billions of money thereby leading to a drop in the economy. Owing to the aforementioned challenges affecting the tourism sector, it is important to predict the estimated future number of tourists visiting Kenya. This prediction will significantly help the Government of Kenya through the ministry of tourism and wildlife (MTW) to plan properly in curbing the draw backs facing the industry.

There are numerous studies that have been done to predict the number of tourists visiting various countries of the world using the SARIMA model. For instance, (Msofe & Mbago, 2019; Makoni et al., 2021; Makoni & Chikobvu, 2018) applied the Box-Jenkins SARIMA model to predict the number of tourists arrival in Zanzibar and Zimbabwe respectively. The models were formulated and fitted to the data using R statistical software. The Information Criterion (AIC) was applied for model selection. The findings of these studies revealed that the future numbers of tourists visits would increase significantly. Therefore, recommending that the recreational facilities be increased to cater for the projected number of tourists expected.

Borhan and Arsad (2018) applied the SARIMA to predict the international tourism demand from the countries of South Korea, Japan and United States of America to Malaysia. The findings revealed that tourists visiting Malaysia from the United States of America and South Korea will continue to increase in the coming days. While a drop was registered in the forecasting from Japan. The study therefore recommended that strategic plans and measures be put in place to able to manage the future increasing tourist arrivals.

Zayat and Sennaroglu (2020) applied the SARIMA and the Holt-Winters models to forecasts international tourist arriving in Turkey. The two models were fitted and compared using the Mean Absolute Percentage Error (MAPE) to determine an effective model for predicting tourist arrivals. SARIMA$(1,1,1)(1,1,1)_{12}$ was identified as the appropriate model for forecasting the Malaysian tourist arrivals with minimum errors than the Holt-Winters model.

Therefore, this study seeks to apply the Seasonal Autoregressive Integrated Moving Average (SARIMA) time series model to forecast future numbers of tourist visits to Kenya. The forecast is done for a period of 3 years from the year 2023 to 2025 after fitting an appropriate model to the data. The information obtained from this study will help the Kenyan government to formulate and implement tourism strategies like enhancing tight security to curb terrorism attacks and promote intertribal unity to embrace tourism in Kenya.

Materials and methods

Data collection and analysis

The secondary data for the number of tourists who visited Kenya in the period between 2011 and 2022 was collected from the ministry of tourism and wildlife. The data can be made available by the corresponding author upon request. The time series data was fitted in R statistical software version 4.1.2 (R Core Team, 2020). In order to select the best SARIMA model, the Akaike Information Criterion technique was applied.

Model Description

Autoregressive (AR) Process

Autoregressive process (AR) of order p is expressed as

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_P X_{t-P} + \epsilon_t \quad (1)$$

X_t denotes the stationary present value that is forecasted at time t, X_{t-1}, $X_{t-2}, \cdots, X_{t-P}$ are the response variables at times $t-1, t-2, \cdots, t-p$ respectively, while $\grave{o}_t$ is the pure random process (Ponziani, 2021). Using a backshift operator B^p, the $AR(p)$ model can be represented by:

$$\phi(B)X_t = (1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_P B^P)X_t \quad (2)$$

Moving average (MA) process

The Moving Average process is denoted by $MA(q)$ for entries up to order q. Moving average models use past errors to forecast present variable.

$$X_t = \beta_0 \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_q \epsilon_{t-q} \quad (3)$$

In the above, q denotes the number of lags in the MA process, $\beta_0, \beta_1, \cdots, \beta_q$ are the parameters to be estimated while $\grave{o}_t$ denotes the white noise which characteristically has a mean of zero and variance σ^2. The MA operator is given by:

$$\theta(B)\epsilon_t = \left(1 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q\right)\epsilon_t \quad (4)$$

ARMA model

The Autoregressive Moving Average process is denoted by ARMA (*p, q)*, the mentioned model is a merger of the simple $MA(q)$ and $AR(p)$ models. The following is a representation of the ARMA (*p, q*) model:

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_q \epsilon_{t-q} \quad (5)$$

Where, $\beta_1 \epsilon_{t-1} + \cdots + \beta_q \epsilon_{t-q}$ is the $MA(q)$, and $\alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p}$ is the $AR(p)$ processes respectively and $\grave{o}_t$ is the random error. The backward shift operator for the ARMA (*p, q*) model is given by

$$X_t = \frac{\theta(B)}{\phi(B)}\epsilon_t \quad (6)$$

Where,

$$\theta(B)\epsilon_t = \left(1 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q\right)\epsilon_t$$

$$\phi(B)X_t = (1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_P B^P)X_t$$

ARIMA model

The Autoregressive Integrated Moving Average (ARIMA) model is a generalization of the Autoregressive (AR) model and the Moving Average (MA) to a differenced data (Montgomery et al., 2015). The ARIMA model was designed by the Box and Jenkins for time series forecasting (Box et al., 2013). The model is denoted by ARIMA(p, d, q) whereby p represents the number of AR terms, d is the number of times the time series is differenced to achieve stationarity, and q denotes the quantity of moving average terms. The ARIMA model uses the past data and reduces it into an AR process which is a recollection of past observations into an integrated process that justifies stationarity (Akuno et al., 2015). The MA utilizes lagged entries of the forecast errors of previous observations to improve present forecast.

SARIMA model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) is an extension of the ARIMA model to a seasonal data (Anastassopoulou et al., 2020; Maleki et al., 2020; Msofe & Mbago, 2019). The model was proposed by Box and Jenkins and is expressed as SARIMA$(p, d, q)(P, D, Q)S$ where, p is the AR order, d is the differenced part, q is the MA

order, P is the SAR order, D is the seasonal differencing and Q is the SMA order. The SARIMA model is given by,

$$\vartheta_P\left(B^S\right)\alpha_p\left(B\right)\left(1-B^S\right)^D\left(1-B\right)^d X_t = \theta_Q\left(B^S\right)\beta_q\left(B\right)z_t \quad (7)$$

Where,

$$\alpha_p(B) = 1-\alpha_1 B-\alpha_2 B^2-\cdots-\alpha_p B^p$$
$$\beta_q(B) = 1-\beta_1 B-\beta_2 B^2-\cdots-\beta_q B^q$$
$$\vartheta_P(B^S) = 1-\vartheta_1 B^S-\vartheta_2 B^{2S}-\cdots-\alpha_P B^{PS}$$
$$\theta_Q(B^S) = 1-\theta_1 B^S-\theta_2 B^{2S}-\cdots-\theta_Q B^{QS}$$

Box-Jenkins methodology

Box and Jenkins proposed the following methodology in time series modeling; model identification, parameter estimation, Diagnostic testing and model selection and forecasting (Box et al., 2013; D. Montgomery et al., 2008).

Model identification

The study attempted to identify an appropriate model structure either, AR, MA or ARIMA. The identification was achieved by observing plots of initial data. A time plot of the time series was essential in determining the need of differencing. If the data happens to be non-stationary, then the study conducted a first order differential to achieve stationarity. Differencing can be iterated until stationarity is achieved with the number of iterations being recorded as *d*. Should we select an ARIMA model then it is denoted as ARIMA (p,d,q). An Augmented Dickey Fuller (ADF) test was conducted to examine the stationarity of the data by investigating the absence or existence of a unit root.

Parameter estimation

In time series analysis, parameters are estimated using the maximum likelihood estimation method and least-square estimation method. In this study, the parameters were estimated using the maximum likelihood estimate because the residuals of the model are assumed to be normally distributed. The Gaussian time series X_t has a covariance matrix denoted as $\Gamma n = E(X_n X'_n)$ where $n = 1, 2, \cdots, n$ (Brockwell & Davis, 2002). Assuming that Γn is a non-linear singular matrix, the likelihood of X_n is given by:

$$\mathrm{L}(\Gamma_\mathrm{n}) = \left(2\pi\right)^{-\frac{\mathrm{n}}{2}}\left(\det\Gamma_\mathrm{n}\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\mathrm{X}_\mathrm{n}^{'}\Gamma_\mathrm{n}^{-1}\mathrm{X}_\mathrm{n}}\right) \quad (8)$$

We then express Γ_n in terms of finite numbers of unknown parameters $\phi_q, \theta_p, \sigma^2$ that is $\Gamma_n = \phi_q, \theta_p, \sigma^2$. A maximum likelihood estimator maximizes the likelihood function $L(\Gamma_n)$ for a selected data set. Given $X_1, X_2, \cdots, X_n$ are identically and independently distributed and n tends to be large enough then the maximum likelihood estimators have a normal distribution with variances that are small as the case of asymptomatic normally distributed estimators. As a result, X_t will not be Gaussian in which case the maximum likelihood estimators become the best estimators. The ARMA models' parameters are numerically calculated by minimizing $\frac{\partial}{\partial\sigma^2} log\left(L(\Gamma n)\right) = \frac{\partial}{\partial\sigma^2} logL\left(\phi_q, \theta_p, \sigma^2\right)$ where θ_p and ϕ_q represent MA and AR coefficients. The initial values of θ and ϕ are fitted from the data as the computer program eventually

systematically search for the values in the reduced log-likelihood function which yields least square estimates. Completion of the iterations, the variance is then computed by maximum likelihood method.

Model selection

The best model has a lower information criterion and the least parameters (Coghlan, 2014). In order to select the parsimonious model, the study will use a stepwise selection criterion utilizing Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and Akaike Information Corrected Criterion (AICc). Letting *Lmax* be the maximum likelihood, k be the number of parameters and n be the sample size then

$$BIC = -2\ln Lmax + klnn \tag{9}$$

Due to their consistent estimates, Bayesian estimates are considered but their difficulty in handling complex models whereby $n < k$ makes the model unsuitable. As such, we may be forced to consider Akaike information criterion (AIC).

$$AIC = 2k - 2\ln Lmax \tag{10}$$

Under the AIC, corrections for the additional parameters are necessary since the principle is based on asymptomatic property. AICc realigns the model parameters resulting in a more parsimonious model (Antonov, 2016). When considering a large sample size, the AICc converges to AIC and as such, the AIC is the preferred criterion by a majority of scholars.

Diagnostic testing

After fitting the model and estimating parameters, the diagnostic test was performed. In time series analysis, diagnostic test is applicable when parameters are estimated using maximum likelihood estimates (Keith & Mcleod, 1994). In this research, diagnostic test was done using residuals. The Q-Q plot and histogram for residual determines if the residuals are independently distributed. The Ljung-Box test checks if the residuals are distributed as a white noise.

Forecasting

Box-Jenkins methodology suggests that the model applied for forecasting has to be stationary and invertible (Dritsakis & Klazoglou, 2019). Once the conditions have been achieved through methods such as ditrending and differencing, the study can then proceed to forecasting.

Results and Discussion

Time series plot

Figure 1: Time series plot for tourists' arrival in Kenya

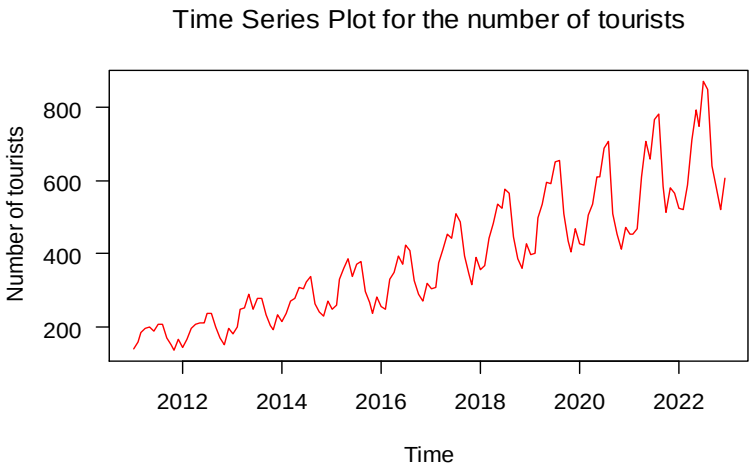
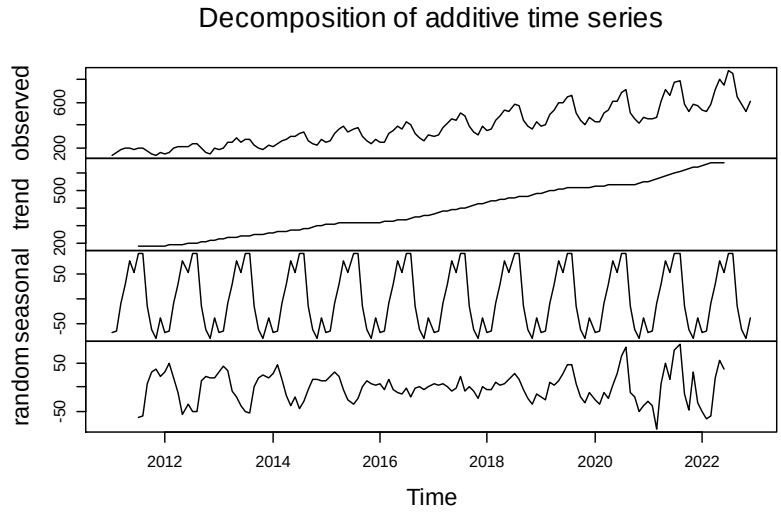


Figure 1 shows the time series plot for the number tourists who visited Kenya from 2008 to 2019. It is observed that the series is non-stationary due to presence of trend in the data. Therefore, either detrending or differencing is required to make the series stationary.

Additive decomposition

Figure 2 shows how the time series data was decomposition additively into different time series components i.e., trend, seasonality, randomness (irregular components) and observed (cyclic variation) as illustrated in Figure 2 below.

Figure 2: Additive decomposition



Differencing and detrending

From Figure 1, it was observed that the series is not stationary due to presence of trend component in the series. In order to remove the non-stationarity component from the series, the series has to either be detrend or differenced. Figure 3 and 4 shows the plot for the differenced and detrend time series data respectively.

Figure 3: Plot of differenced data

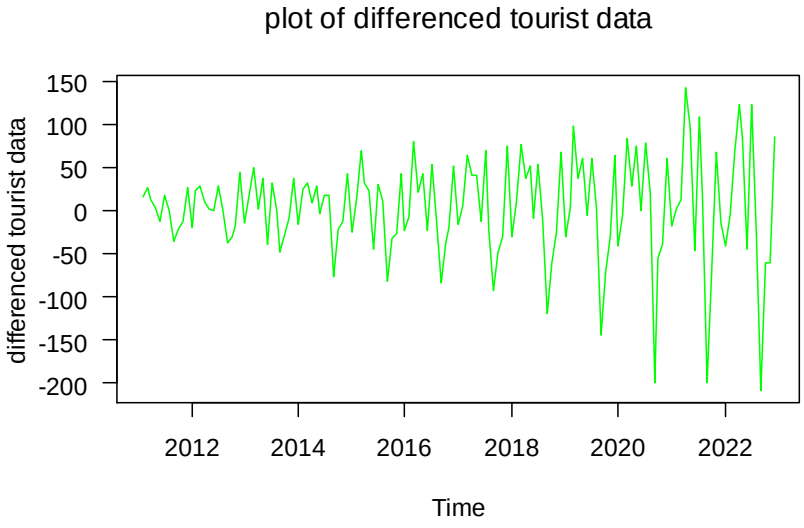
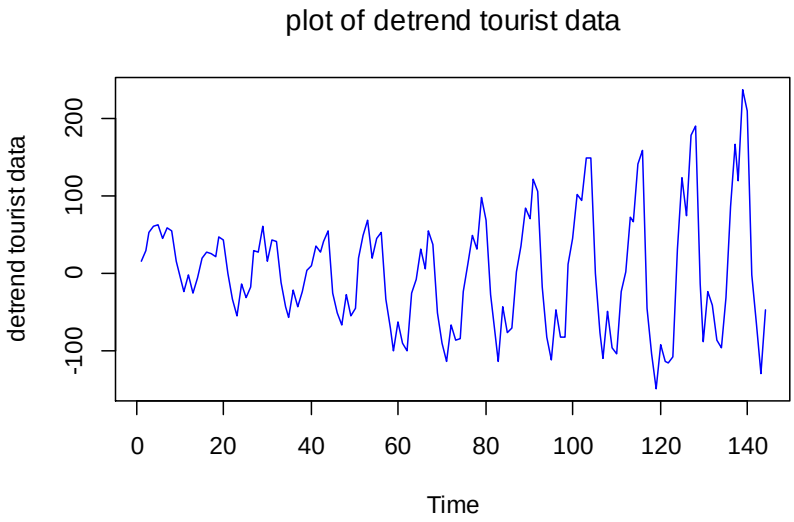
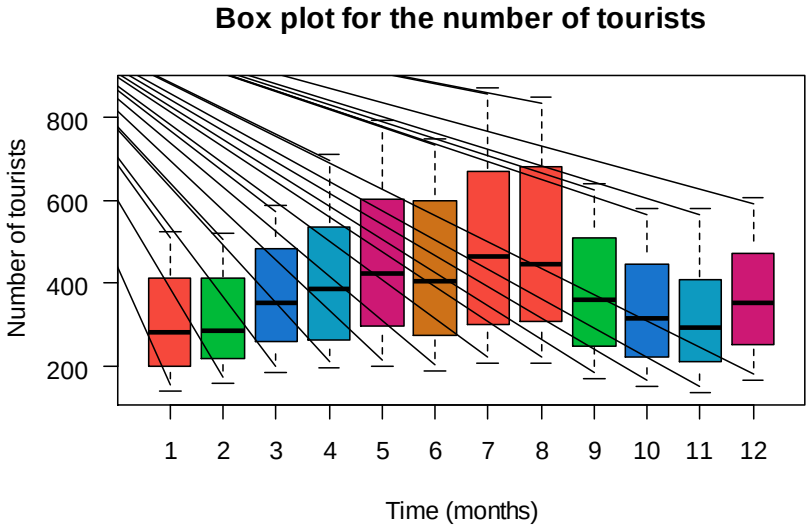


Figure 4: Plot of detrend data



From Figure 3 and 4, it is evident that both differencing and ditrending removes the non-stationarity component from the series. Figure 5 shows the boxplot of the number of tourists that visited Kenya between 2011-2023. From the boxplot, it is observed that the months of July and August have the highest number of tourists.

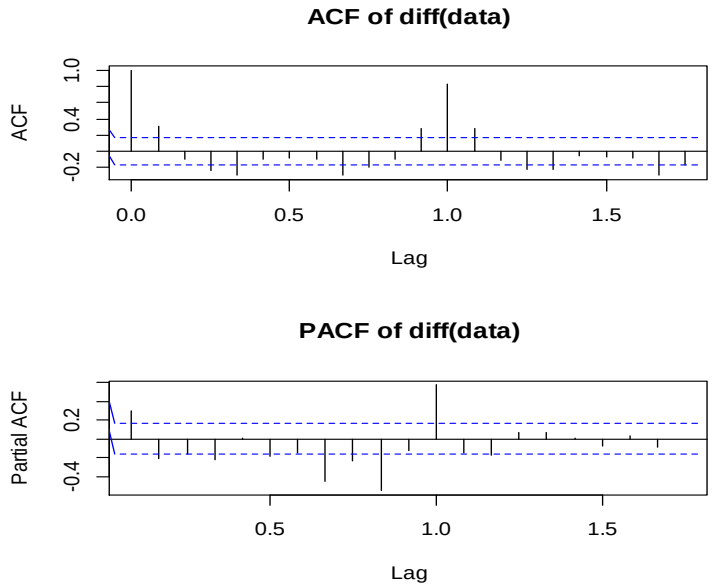
Figure 5: Boxplot for tourists



Model identification

The Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) are plotted to determine the Autoregressive (AR) order and the Moving Average (MA) order respectively. From the ACF and PACF plot in Figure 6, the spikes lie outside the significant zone implying the residuals are not random. Therefore, more information is needed to estimate the moving average (MA) order and the autoregressive (AR) order.

Figure 6: ACF & PACF for the differenced data



Model selection and parameter estimation

Since the ACF and PACF did not provide enough information to estimate the MA order and the AR order, the automated ARIMA function was used in R software to generate different models and select the perfect model with lower AIC value. After fitting the model to the data, $\text{ARIMA}(2,1,1)(0,1,0)_{12}$ was selected as the best model with AIC value of 1017.85. The estimated variance of the model ($\hat{\sigma}^2$) was 132.3. Table 1 shows the regression coefficient for z test. All the coefficients are significant at 5% confidence level.

Table 1: Z test of the coefficients

Coefficients	Estimate	Std. Error	z value	p-value
ar1	0.596	0.089	6.710	0.00
ar2	0.214	0.088	2.436	0.02
ma1	-0.98	0.029	-33.62	0.00

Diagnostic checking

The residuals were checked to see if they are randomly and identically distributed. The test was done using the normal probability (Q-Q) plot, histogram for residuals and the Ljung Box test. From Figures 7 and 8, the normal Q-Q plot and the bell-shaped histogram for the residuals implies that the model residuals are normally and identically distributed.

Figure 7: Normal Q-Q plot of residuals

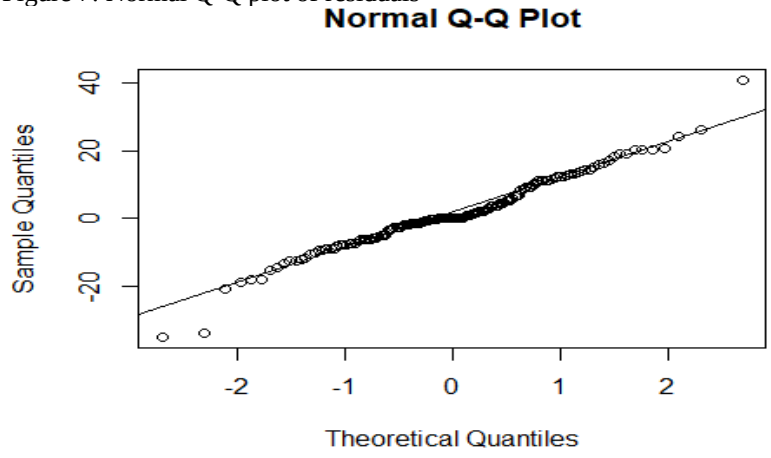


Figure 8: Histogram of residuals

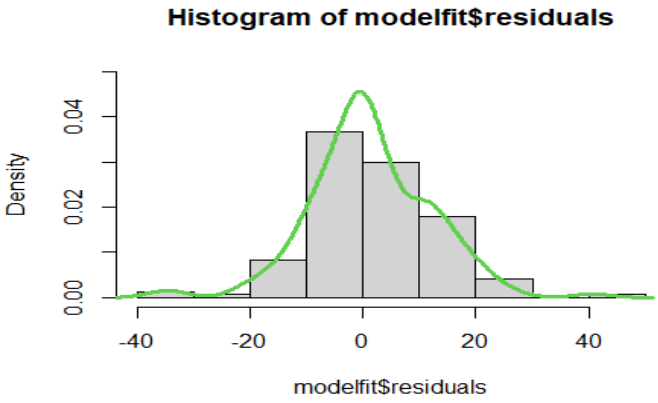


Table 2 shows the Box-Ljung test statistics. The higher p-value for the modified Ljung test implies non-significance therefore, we fail to reject the null hypothesis which states that the residuals are independently distributed and conclude that there is no serial correlation between lags. Therefore, the model is distributed as white noise.

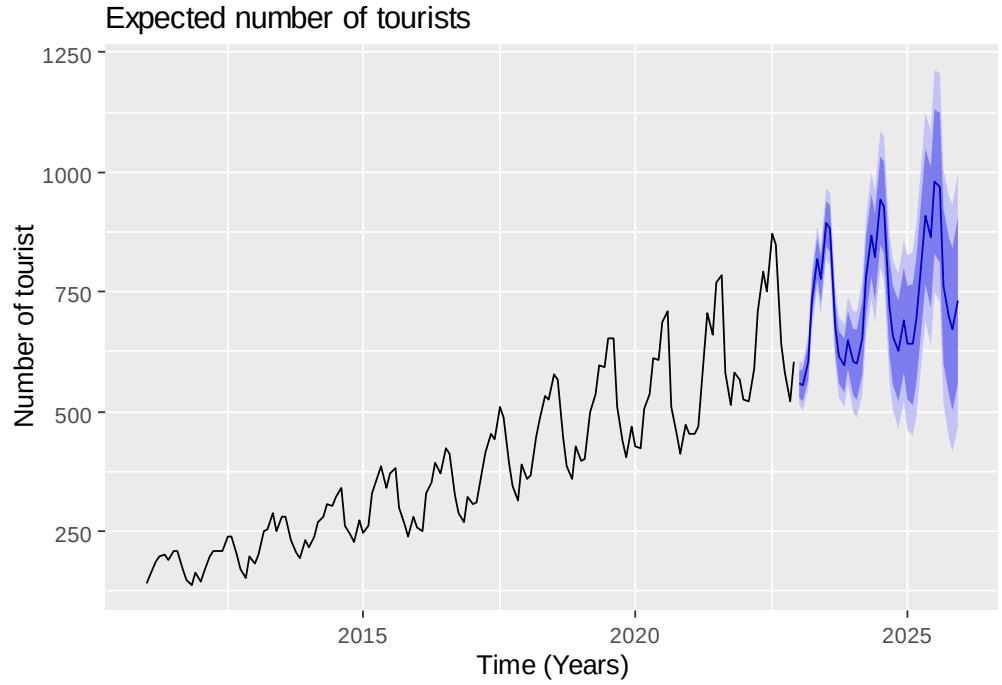
Table 2: Ljung Box test

$\chi-$squared	lags	p-value
20.153	20	0.4484

SARIMA prediction

The future tourist prediction in Kenya was done at a 95% significance level using the best-selected model and, the prediction was done for a period of three years from 2023 to 2025. Figure 8 shows the prediction plot for the number of tourists expected to visit Kenya between 2023 and 2025. The plot suggests that the expected number of tourists will increase significantly for the next three years.

Figure 9: Tourists prediction



Conclusion

The study adopted the Seasonal Autoregressive Integrated Moving Average time series analysis technique to predict the number of tourists expected to visit Kenya between 2023 and 2025. We used the secondary data for the number of tourists who visited Kenya from 2011 to 2022 as recorded by the ministry of tourism and wildlife. The model was formulated and fitted to the data using R statistical software and the $\mathrm{ARIMA}(2,1,1)(0,1,0)_{12}$ was identified as the best prediction based on the information criterion. The model passed the diagnostic test and was applied for prediction. Forecasting of the number of tourists arriving to Kenya was done at 95% confidence level for a period of four years (2023-2025) using the fitted model. From the prediction plot, the expected number of tourists were observed to increase significantly. The study recommends that the recreational facilities and accommodations should be maintained and increased to cater for the projected numbers of tourists expected to visit Kenya in 2023/2025 span. The study also recommends that the government should strategize on how to beef up security to curb terrorism attacks and tribal conflicts in order to attract more tourists.

References

Akuno, A. O., Otieno, M.O., Mwangi, C.W. & Bichanga, L.A. (2015). Statistical Models for Forecasting Tourists' Arrival in Kenya. *Open Journal of Statistics*, 5(1), 60–65.

Anastassopoulou, C., Russo, L., Tsakris A. & Siettos, C. (2020). Data-Based Analysis, Modelling and Forecasting of the COVID-19 outbreak. PLoS ONE 15(3): e0230405

Antonov, A. (2016). Economics and Political Economy. *Automating Analytics: Forecasting Time Series in Economics and Business*, 3(2).

Borhan, N. & Arsad, Z. (2018) 'Forecasting International Tourism Demand from the US, Japan and South Korea to Malaysia: A SARIMA Approach, AIP Conference Proceedings, 1605(1).

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (2013) *Time Series Analysis: Forecasting and Control*. Fourth edition. New York: Wiley.

Brockwell, P. J. & Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Second Edition. New York: Springer.

Coghlan, A. (2014). Using R for Time Series Analysis-Time Series 0.2 Documentation. Available at: http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/src/timeseries.html#decomposing-seasonal-data.

Dritsakis, N. & Klazoglou, P. (2019). Time Series Analysis Using ARIMA Models: An Approach to Forecasting Health. *International Econmics*, 72(1), 77–106.

Keith, W. & Mcleod, A. (1994). Chapter 7 Diagnostic Checking. *Development in Water Science*, 45, 235–253.

Makau, J., Njuru, S. & Ocharo, K. (2018). Macroeconomic Environment and Public Debt in Kenya. *International Journal of Economics*, 3(1), 49–70.

Makoni, T. & Chikobvu, D. (2018). Modelling and Forecasting Zimbabwe's Tourist Arrivals Using Time Series Method: A Case Study of Victoria Falls Rainforest. *Southern African Business Review*, 22(1), 1-22.

Makoni, T., Chikobvu, D. & Sigauke, C. (2021). Hierarchical Forecasting of the Zimbabwe International Tourist Arrivals. *Statistics, Optimization & Information Computing* 9(1), 137–156.

Maleki, M., Mahmoudi, M.R., Wraith, D. & Pho, K-H. (2020). Time Series Modelling to Forecast the Confirmed and Recovered Cases of COVID-19. *Travel Medicine and Infectious Disease*, 37, 1-6.

Montgomery, D. C., Jennings, C. L. & Kulahci, M. (2015). Time Series Analysis and Forecasting. In *Introduction to Time Series Analysis and Forecasting*. Second edition,

pp. 1–671. Edited by Jennings, C.L., Montgomery, D. & Kulahci, M. New York: Wiley.

Msofe Z. A. & Mbago, M. C. (2019). Forecasting International Tourist Arrivals in Zanzibar Using Box – Jenkins SARIMA Model. *General Letters in Mathematics*, 7(2), 100–107.

Ponziani, R. M. (2021). The Inflation Forecasting of Major Cities in East Kaliman- tan: A Comparison of Holt-Winters And SARIMA Models. *International Journal of Data Science Engineering, and Analytics*, 1(2), 1–11.

R Core Team (2020). R: A Language and Environment for Statistical Computing. Available at: https://www.r-project.org/.

Zayat, W. & Sennaroglu, B. (2020). Performance Comparison of Holt-Winters and SARIMA Models for Tourism Forecasting in Turkey. *Doğuş Üniversitesi Dergisi*, 21(2), 63–77.